(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 July 2003 (10.07.2003)

PCT

(10) International Publication Number
WO 03/056454 A1

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/EP02/14266

(22) International Filing Date:
14 December 2002 (14.12.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
01131036.4 28 December 2001 (28.12.2001) EP

(71) Applicant *(for all designated States except US)*: **THOMSON LICENSING S.A.** [FR/FR]; 46 Quai A. le Gallo, F-92100 Boulogne-Billancourt (FR).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **WINTER, Marco** [DE/DE]; Böhmerstr. 17, 30173 Hannover (DE). **ADOLPH, Dirk** [DE/DE]; Wallbrink 2, 30952 Ronnenberg (DE). **HÖRENTRUP, Jobst** [DE/DE]; Vossstr. 35, 30161 Hannover (DE).

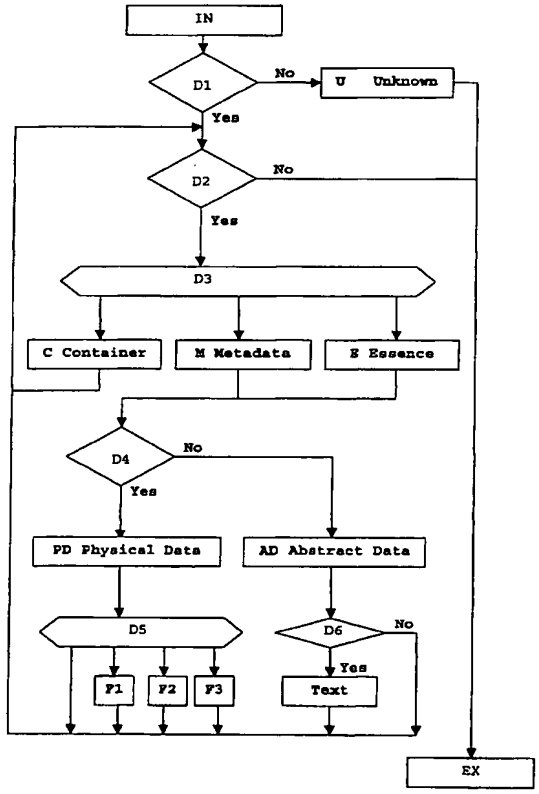
(74) Agent: **RITTNER, Karsten**; European Patent Operations, Karl-Wiechert-Allee 74, 30625 Hannover (DE).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: METHOD AND APPARATUS FOR AUTOMATIC DETECTION OF DATA TYPES FOR DATA TYPE DEPENDENT PROCESSING



(57) Abstract: A method for automatic detection of data types for data type dependent processing has two orthogonal classification systems defined, and determines for incoming data items a data type according to the first classification system and another data type according to the second classification system. The first classification system comprises the data types Essence (E), Metadata (M) and Container (C). The second classification system comprises the data types Physical Data (PD) and Abstract Data (AD). A default data type may be defined for data items not being uniquely classifiable. Advantageously, the inventive method can be used when different classes of data items require different methods for processing, e.g. content searching.

WO 03/056454 A1

European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Published:

— *with international search report*

Method and apparatus for automatic detection of data types for data type dependent processing

The invention relates to a method and an apparatus for the classification, organization and structuring of different types of data, which can be used e.g. for data sorting, data storage or data retrieval.

Background

The capacity of digital storage media like hard disks or rewritable optical disks for personal recording of video and other data grows continuously. This results in new concepts like e.g. the so-called home server, which is a central storage device with large capacity for recording any kind of data within the home. Such applications also require new ways to organize the recorded data, search for content and access specific recordings.

For this purpose data about data, often referred to as metadata, can be used. Various industry groups and standard bodies have been developing metadata standards for different purposes and applications. In multimedia applications, metadata typically are data about audiovisual (AV) data, these AV data often being called 'essence'. However, a Data Base Management System (DBMS) that shall be able to handle data of various data types correctly requires a definition of data types, and a method to distinguish between them.

Invention

The invention is based on the recognition of the facts described in the following:
In devices providing a DBMS for handling of incoming data,

including incoming metadata, it is necessary to classify said incoming data, and especially incoming metadata, since different processing is necessary for different kinds of metadata. For example, a text query is not suitable for metadata containing a picture in the well-known Graphics Interchange Format (GIF).

The problem to be solved by the invention is to classify the data automatically, such that a DBMS can utilize the result of the classification for correct data handling. This problem is solved by the method disclosed in claim 1 and by the apparatus disclosed in claim 5. The output of such apparatus may be directed towards e.g. a DBMS.

According to the invention, Metadata can be defined as data sets consisting of two parts, namely a first part being a link, the link pointing to a reference data set, and a second part being any data referring to said link. In the following, said first part is referred to as MD_LINK, and said second part is referred to as MD_LOAD. Any data item that does not contain at least one MD_LINK and a related MD_LOAD is defined to be Essence. Metadata often occur together with other Metadata or Essence, combined in a logical entity like e.g. a file on a hard disc. Such mixture of different kinds of Essence and Metadata is in the following called 'Container'. Popular examples for such Containers are Hypertext Markup Language (HTML) files, or Portable Document Format (PDF) files.

Further, according to the invention there is another type of classification possible. Data may require interpretation through the device before they can be used. In this case the data are defined to be Physical Data, if the device has a method for interpretation defined, otherwise Abstract Data. If e.g. a picture is stored in GIF format, and the device can interpret GIF format and display it as a picture, it is

classified as Physical Data. If the device cannot interpret GIF format, the data is classified as Abstract Data. Further examples for Abstract Data are text files, and other files that cannot be interpreted through the device.

The previously defined two types of classification are not exclusive, but complementing each other. Further, the described classification of data is not absolute, but system dependent, and therefore only locally relevant.

Advantageously, this classification allows the device to handle different data types correctly, differ between Metadata, Essence, Container, Physical Data and Abstract Data, and thus permit a generalized access method upon said data types. With this knowledge, the device can decide e.g. which type of data-query to use, how to interpret data, and if some data can be disregarded for a certain query.

Advantageous additional embodiments of the invention are disclosed in the following text, and in the respective dependent claims.

Drawings

Exemplary embodiments of the invention are described with reference to the accompanying drawings, which show in:

Fig. 1 the two systems, or dimensions, of data classification;

Fig. 2 an example for a Container containing Essence and Metadata;

Fig. 3 an example for Abstract Metadata;

Fig. 4 an example for Physical Metadata; and

Fig. 5 an exemplary flow-chart for the method according to the invention.

Exemplary embodiments

According to the invention, the two types, or systems, of classification can be understood as two dimensions, as shown in Figure 1. A data item may either be Essence E or Metadata M, and either Physical Data PD or Abstract Data AD. Therefore the possible data types are Physical Essence PE, Physical Metadata PM, Abstract Essence AE or Abstract Metadata AM. Further, a data item may also be a Container C, if it contains other data items.

The classification of data is not absolute, but subjective from the sight of the device, and therefore only relevant within a system, e.g. DBMS. It may happen that e.g. one system can interpret a link while another system cannot interpret the same link. Therefore it may happen that e.g. one system classifies certain data as Metadata, consisting of MD_LOAD and MD_LINK, while another system classifies the same data as Essence because it cannot interpret the link. Another example is that e.g. one system can reproduce an MPEG audio layer 3, or MP3, coded file, while another system cannot interpret the MP3 format. In this case the first system classifies an MP3 coded file as Physical Data, but the second system classifies the same file as Abstract Data.

Text is to be regarded as Abstract Data, because text is always a format for saving data. Formatted text can represent a direct physical representation of data, e.g. the PDF format. The format information represents only support information, i.e. if format information is extracted from a

PDF file, the pure text being the main information will remain. If the text is extracted, the main information will be lost. Due to the fact that the text represents the main information, also formatted text will be regarded as Abstract Data.

A device as disclosed in claim 5 will execute the following procedure when receiving data on its input:

If the data contain more than one data item, the output may be: "Data is a Container". More details are given below. Classification may stop here, or may be extended to some, or all, leaves of the hierarchically structured data tree within the Container.

If data are Metadata, the output may be: "Data are Metadata".

Otherwise the output may be "Data are Essence".

If data are Physical Data, an additional output may be "Data are Physical Data".

Otherwise, if data are Abstract Data, an additional output may be "Data are Abstract Data".

Advantageously the device can detect and output the type of Physical Data, e.g. "Data is a color picture (24bit) with the resolution x=200 pixels and y=400 pixels".

If the data format is unknown to the device, and therefore the device is not able to classify the data as Container, Metadata, Essence, Abstract Data or Physical Data, the output may be any default-type output, e.g. "Data type is unknown" or "Data are Essence and Abstract Data".

Additionally, it is helpful if the device detects whether data is text or not:

If data is Abstract Data and text, the output may be additionally "Data is Text".

This may be implemented by searching for known words, e.g.

from an electronic dictionary, or searching for groups of characters separated by blanks.

If the input data is a Container, an additional output may be "Data is a Container, i.e. more metadata or essence are contained". Optionally, precise details can be included: "The Container CONTAINS at least 1 Metadata and 1 Essence", or "The Container CONTAINS no Metadata at all" or even "The Container CONTAINS exactly N Metadata items", with N being the amount of Metadata contained in the Container.

If the device can detect the format of the analyzed data, it may output it additionally: "Data format is X". 'X' is the format. Examples for 'X' can be e.g. 'HTML' or 'JPEG'.

Figure 2 shows an example for a data file containing a combination of Essence and Metadata in the well-known HTML format. In the following, the classification of all elements according to the invention is described.

First the device detects that the first line is <html>, and that therefore the data file should be HTML formatted. It is assumed that the device can interpret the HTML format, and therefore interprets items with "href" attributes in HTML files as links. Since HTML formatted files usually contain a hierarchical structure, the leaf elements of the hierarchy tree are analyzed first. The first element from Fig.2

<title>This is the title</title>

is classified as Essence because there is no link attached to the element.

The element

W3C HOME

is classified as Metadata, with the string "W3C HOME" being the Essence, or MD_LOAD, and the string "href=http://w3c.org" being the related link, or MD_LINK.

The next leaf element
This is a paragraph
contains no link and is therefore classified as Essence.

The next leaf element
<img src="image.gif">
is also classified as Essence because it is only a link, i.e. it contains no MD_LINK with related MD_LOAD. Therefore it cannot be Metadata. The purpose of this link is to reference further Essence, namely the picture data.

When all elements of the first level of hierarchy are analyzed, the next level is investigated. The element
<head>
<title>This is the title</title>
</head>
is classified as Essence because it contains no link, but only one element, the element being Essence.

The element

<img src="image.gif">

is classified as Metadata, with <img src="image.gif"> being the MD_LOAD part and the "href" attribute being the related link.

The next element
<body>
...
</body>
is classified as Container because it groups together Metadata items and Essence items.

Finally, the element

<html>
...
</html>

is also classified as Container. It groups together an Essence element, namely the <head> element, and a Container, namely the <body> element.

Figure 3 shows an example for Abstract Metadata. Several data items 3R,3M are grouped in a data unit 3C. The data unit 3C could be e.g. an HTML file. For one of said data items the device has detected that it contains a link 3L, symbolized by the cursor switching from an arrow to a hand when pointing to the text 3E. Since the text 3E and the link 3L belong together, and the text 3E is Essence, they form a Metadata item 3M, and the link 3L is a Metadata link pointing to a reference 3REF outside the data unit 3C. Since the Essence 3E of the Metadata item 3M is text, and text is Abstract Data, the Metadata item 3M is an Abstract Metadata item. Remaining data items 3R within the data unit 3C are any text and a picture. The data unit 3C is a Container, since it contains at least one Metadata item 3M and other, remaining data items 3R.

Figure 4 shows an example for Physical Metadata. Several data items 4R,4M are contained in a data unit 4C, the unit 4C being e.g. an HTML file. In this case, the device has detected that the picture 4E is associated to a link 4L, symbolized by the cursor switching from an arrow to a hand. The link 4L is pointing to a reference 4REF outside the data unit 4C. Since the picture 4E and the link 4L belong together, they form a Metadata item 4M, with the picture 4E being the Essence of this Metadata. Said Essence 4E is e.g. a JPEG formatted picture, and in the HTML file it may be referenced e.g. as <img src=Anton.jpg width=108 height=73>. Since the device can display it, it is Physical Data, and the Metadata item 4M is Physical Metadata. The data unit 4C

is a container, because it contains at least one Metadata item 4M and other items 4R.

Figure 5 shows an exemplary flow chart of the inventive method. The purpose of the invention is to classify different types of incoming data IN. The incoming data IN are being analyzed, and a first decision block D1 decides whether the format of the incoming data can be detected. If not, 'Unknown' is indicated as an output, and the classification finishes at an end state EX. If the format is known, e.g. HTML, then a second decision block D2 may decide if the incoming data contains unclassified elements. If the answer is 'Yes', the next unclassified data item is picked and forwarded to a third decision block D3. This decision block D3 may decide whether said data item is a Container C, Metadata M or Essence E. The decision is 'Container' if the data item contains another data item already classified as Metadata. The decision is 'Metadata' if the data item contains a link with essence relating to that link. In all other cases the decision is 'Essence'. The decision made in the third decision block D3 is indicated at the output. If the analyzed data item is a Container C, then the procedure returns to the second decision block D2 again, otherwise a fourth decision block D4 is entered. Said fourth decision block D4 decides whether the device can interpret the data item, such that it may disclose further information to the user, e.g. a displayable picture. If the answer is 'Yes', it is indicated at the output that said data item is Physical Data PD, otherwise Abstract Data AD. In the case of said data item being Physical Data PD, format detection may have been done implicitly in said fourth decision block D4. Then a fifth decision block D5 may detect format details and decide whether the detected format shall be indicated, and if so, the format F1,...,F3 may be indicated at the output. In the case of said data item being Abstract Data AD, a sixth decision block D6 may decide if the data contains text. If

so, this is indicated at the output. If the data item is Abstract Data AD and not text, no further indication is generated. Then the procedure is repeated from the second decision block D2 that decides if further unclassified elements are contained. If this is not the case, then the data item has been classified completely and the end state EX is entered. This embodiment of the invention analyzes all hierarchy levels and leaf elements of Containers, but other embodiments may analyze only some hierarchy levels or leaf elements of Containers.

Advantageously, the described method for data classification can be used in devices for data sorting, data storage e.g. DBMS, or data retrieval e.g. browsers. The described method can be used when different classes of data require different processing, e.g. different search algorithms, different storage methods or areas, different compression methods or different presentation methods.

The invention can be implemented in a separate device, which will classify incoming data with respect to its format, content, and relation to other data, e.g. link, and which provides information about data. This information is especially necessary when it is to recognize, whether these data contain links or these data need special query-methods.

The device can be part of another device or can be realized as hardware or software, e.g. as an application or a plug-in in a PC. Further, it can be updated, e.g. via the Internet or via other sources, so that more and more formats can be recognized, thus this device will update itself and get more and more efficient.

Claims

1. Method for automatic detection of data types for data type dependent processing by a technical device, **characterized in**
   a) receiving data (IN) of different data types,
   b) analyzing said received data,
   c) detecting (D1) the format of the received data,
   d) using said detected format for evaluating (D3) whether said data contain
      - at least one machine-interpretable link and associated data (M),
      - any other data (E), e.g. text, picture data, links, except data of said first type (M), or
      - a mixture of said machine-interpretable link and associated data (M) with said other data (E),
   e) evaluating (D4) whether said technical device is able to interpret said data for reproducing a physical representation of said data, and
   f) supplying the result (M,E,C) of said first evaluation and the result (PD,AD) of said second evaluation to a device or process for data type dependent processing of said data (IN).

2. Method according to claim 1, wherein for data being interpretable (PD) by said technical device is also indicated whether the format type of said data is one of a number of specified format types (F1,...,F3).

3. Method according to any of claims 1-2, wherein for data being not interpretable (AD) by said technical device is also indicated if it is text.

4. Method according to any of claims 1-3, wherein said technical device is a data sorting device, a database management system or a data content browser.

5. Apparatus for automatic detection of data types for data type dependent processing **characterized in** that the method according to any of claims 1-4 are used.

5

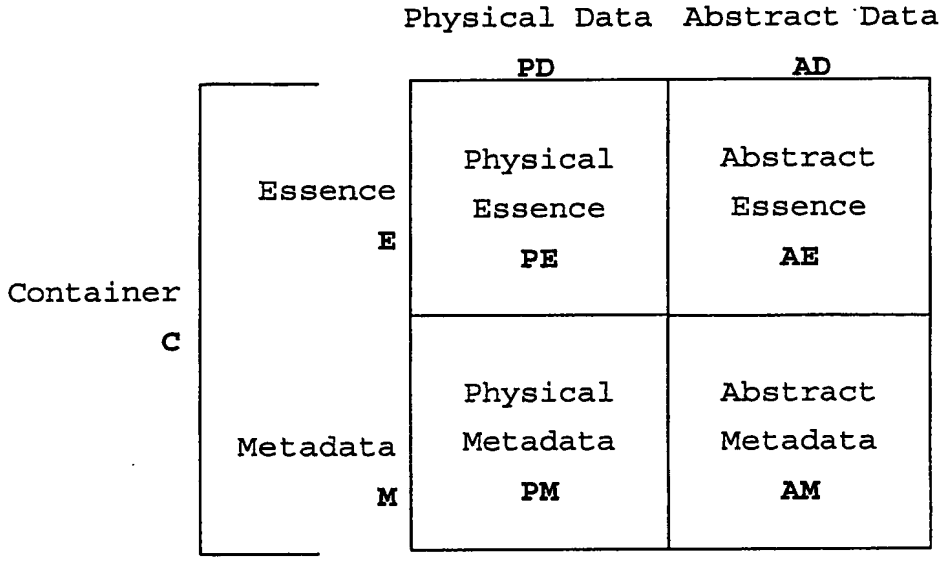


Fig. 1

```
<html>
   <head>
      <title>This is the title </title>
   </head>
   <body>
      <a href="http://www.w3c.org">W3C HOME</a>
      <p> This is a pararaph </p>
      <a href="http://www.w3c.org">
         <img src="image.gif">
      </a>
   </body>
</html>
```

Fig. 2

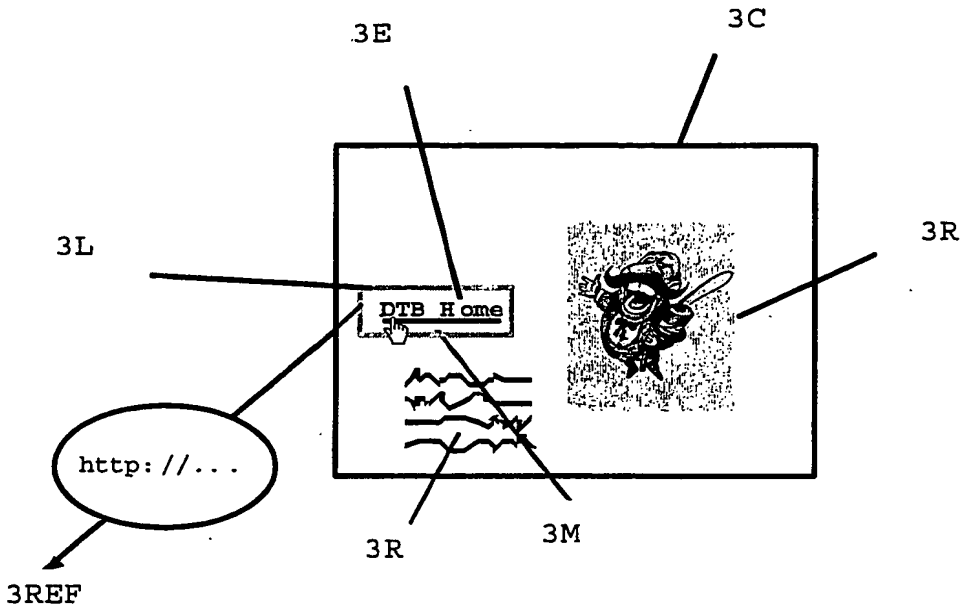


Fig. 3

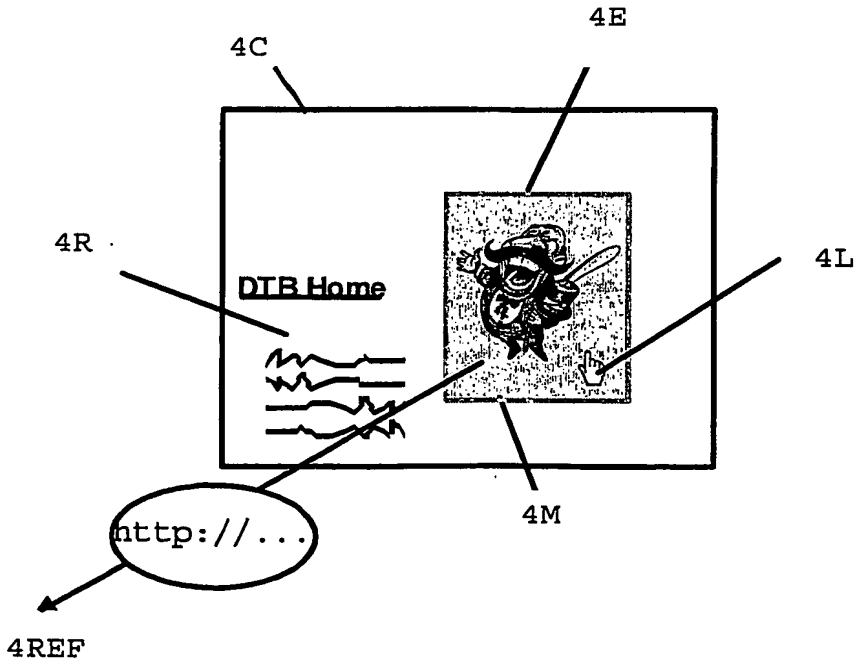


Fig. 4

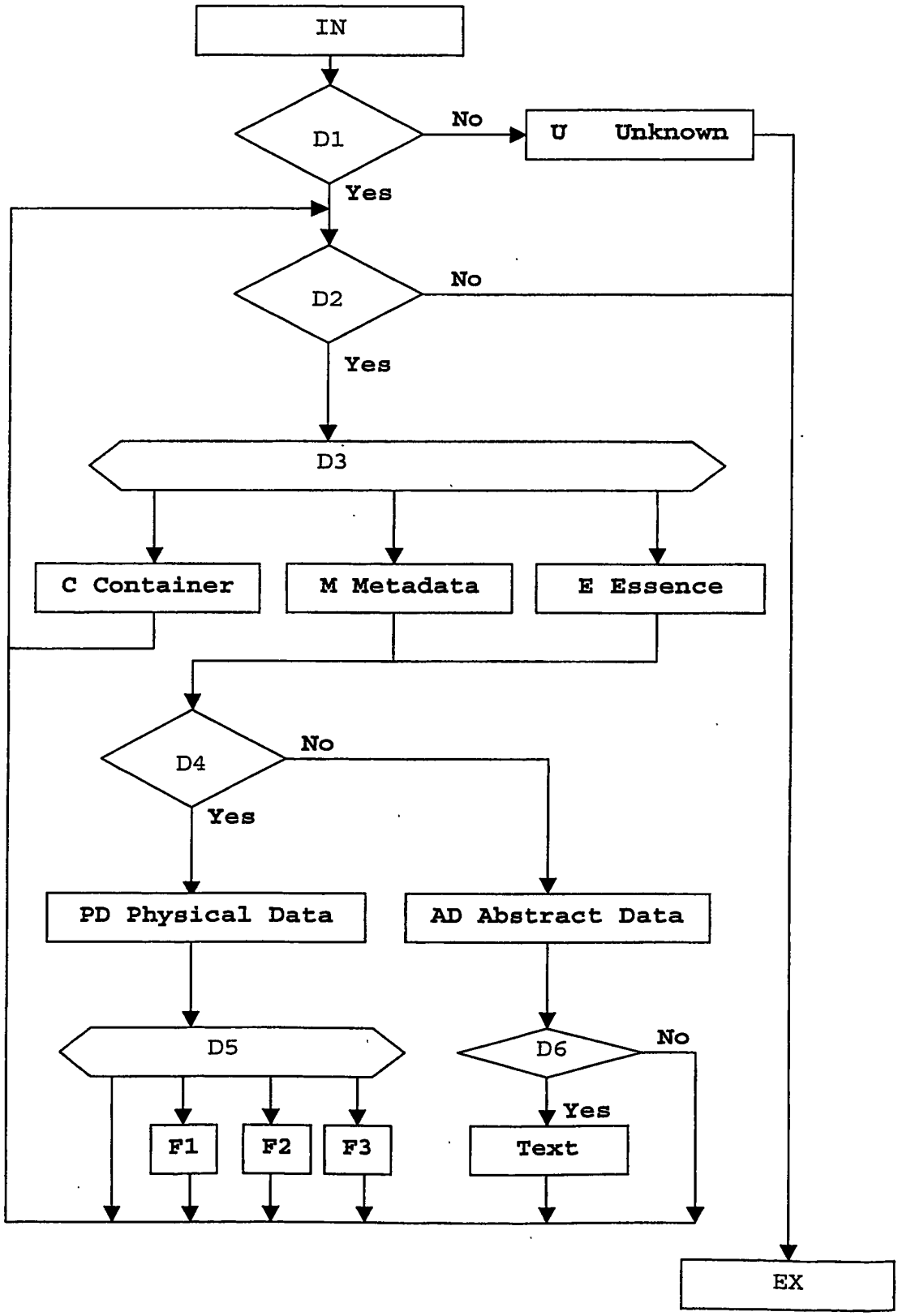

IN
D1
No
U Unknown
Yes
D2
No
Yes
D3
C Container
M Metadata
E Essence
D4
No
Yes
PD Physical Data
AD Abstract Data
D5
D6
No
Yes
F1
F2
F3
Text
EX

Fig. 5

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP 02/14266

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, IBM-TDB

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SHOENS K ET AL: "The Rufus system: information organization for semi-structured data" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, XX, XX, 24 August 1993 (1993-08-24), pages 97-107, XP002151011 page 98, left-hand column page 100 -page 101 page 102, right-hand column, paragraph 3 page 104, paragraph 3 ---	1-5
A	US 5 864 870 A (GUCK RANDAL LEE) 26 January 1999 (1999-01-26) column 1, line 59 -column 2, line 25 column 3, line 24-55 ---	1-5

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
31 March 2003	07/04/2003

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Correia Martins, F

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP 02/14266

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GB 2 212 636 A (AMOCO CORP) 26 July 1989 (1989-07-26) page 2, paragraph 2; claims 1-7	1-5
A	"Unix 'file' command" UNIX MANUAL, 'Online! XP002216604 Retrieved from the Internet: <URL:http://www.rt.com/man/file.1.html> 'retrieved on 2002-10-14! the whole document	1-5

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/EP 02/14266

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5864870 A	26-01-1999	NONE	
GB 2212636 A	26-07-1989	NONE	